DATAFUSION®

From Authority Files to Ontologies: Knowledge Management in a Networked Environment

Joseph A. Busch

September 29, 1999

DATAFUSION, Inc. ©1999

Topics

DATAFUSION®

- 3000 years of library science.
- Infomediation and eCommerce.
- Controlled vocabularies.
- Solutions.

DATAFUSION, Inc. 1999

3000 years of library science

DATAFUSION®

1200 BC
Clay tablets
Papyrus scrolls

400 BC
Library at Alexandria

200 BC
Qin Dynasty Imperial Library

300
Roman private & public libraries

700
Bunko literary storehouses
Parchment codices

... and information technology

DATAFUSION, Inc. 1999

3000 years of library science

DATAFUSION®

1000's
Movable type
Monasteries
Universities

1300's
Libraries in Europe

1400's
Printing press
Imperial Library

1600's
Bodleian Library Harvard University Library

1800's
Library of Congress Boston Public Carnegie libraries Dewey Decimal Classification

... and information technology

DATAFUSION, Inc. 1999

3000 years of library science

DATAFUSION®

1900-1920

Cutter's *Principles*
Ranganathan's *Prolegomena*
Bookmobile

1920-1940

Electronic mass media (radio)
Paperbacks

1940-1960

Digital computing
TV mass media
Cryptography
UDC
NLM

1960-1980

Text searching
OCLC & RLG
IR

1980-2000

Personal computing
Internet mass media
Search engines
Digital libraries
eCommerce
Portals
UMLS
eMail

... and information technology

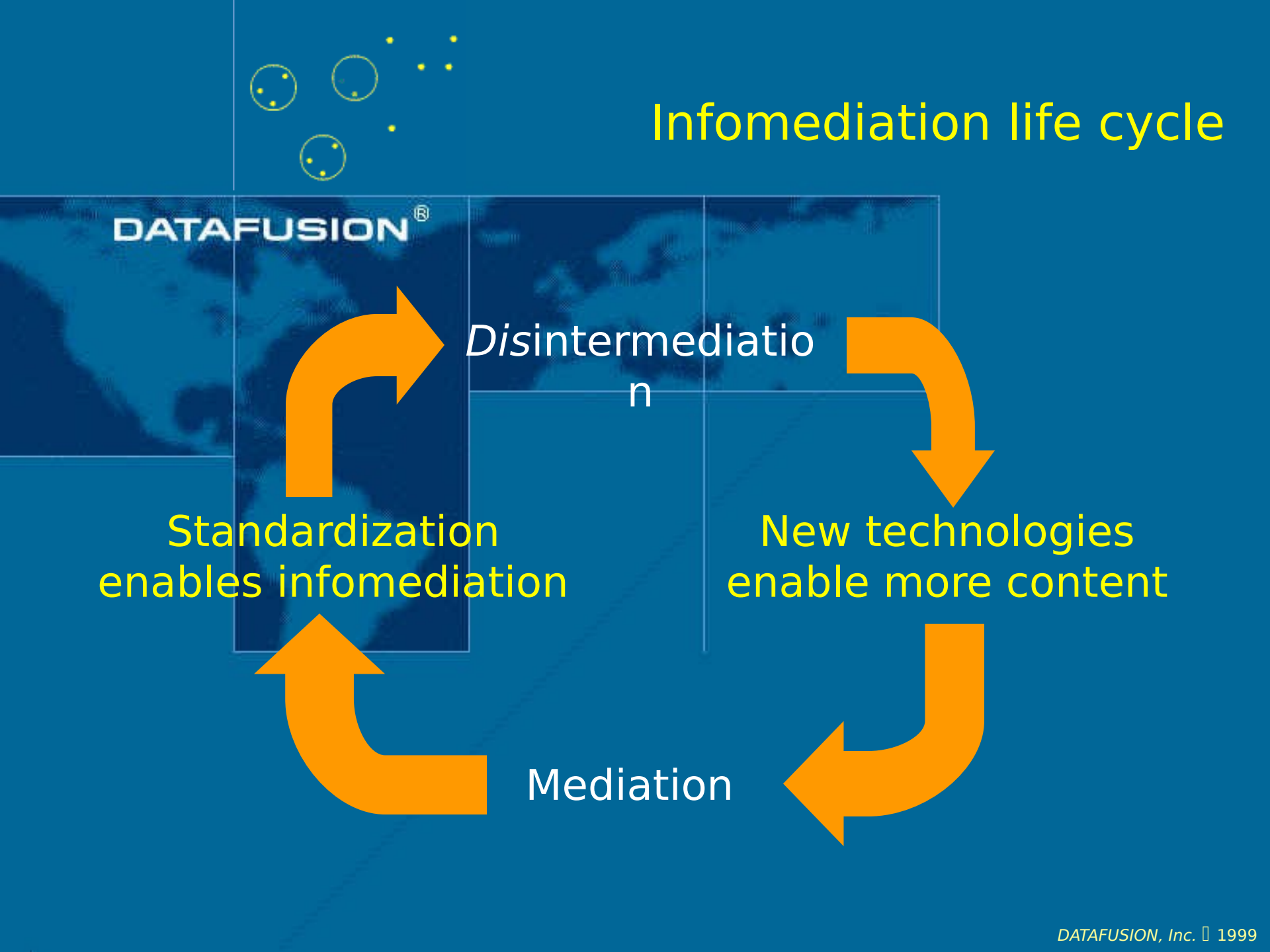
DATAFUSION, Inc. © 1999

DATAFUSION®

- 3000 years of library science.
- Infomediation and eCommerce.
- Controlled vocabularies.
- Solutions.

DATAFUSION, Inc. 1999

Infomediation life cycle

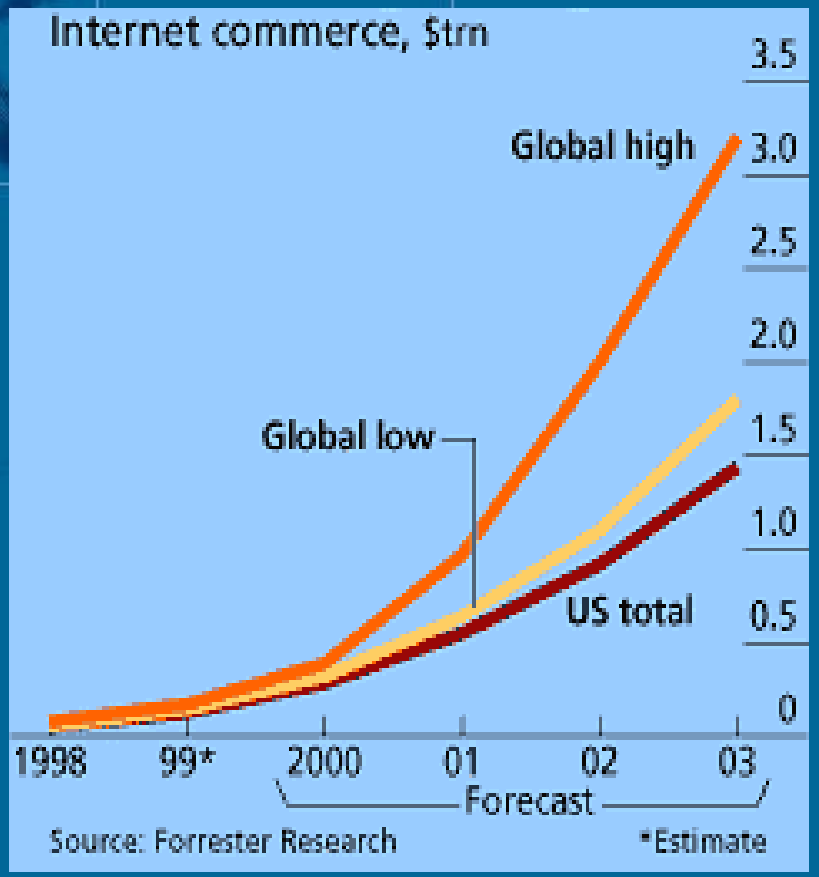


DATAFUSION, Inc. © 1999

Rise of Internet commerce

DATAFUSION®

- Advertising placement
- Consumer shopping
- Consumer auctions
- Pay-per-view content
- Business-to-business marketplace



DATAFUSION, Inc. 1999

Why controlled vocabularies are important

DATAFUSION®

There has to be some agreement on definitions to ensure that there is a shared language of business on the Internet.

The Economist Survey of Business and the Internet

(June 26, 1999)

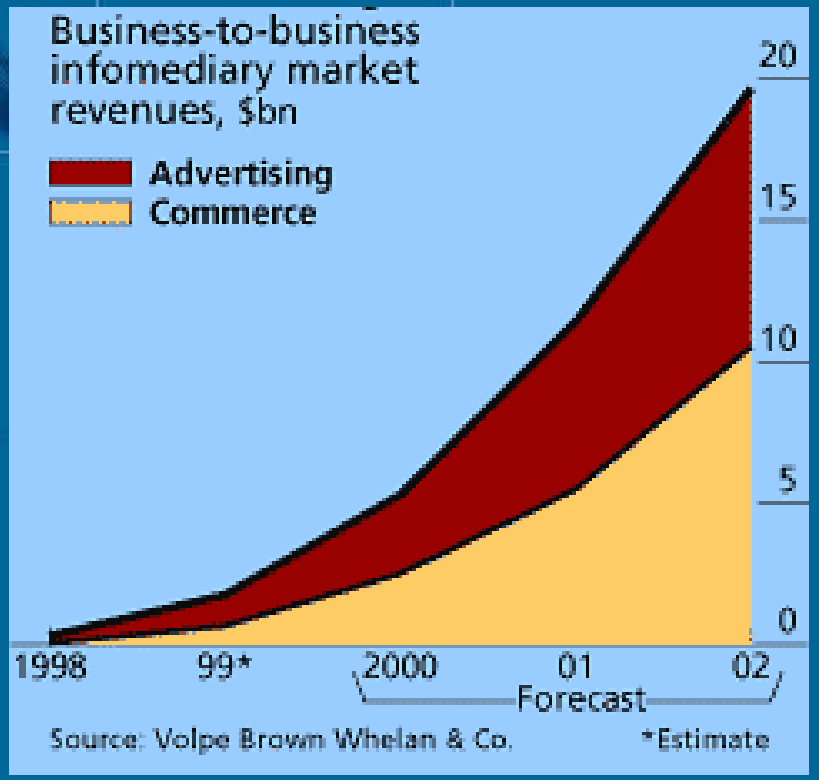
DATAFUSION, Inc. 1999

Rise of infomediation

DATAFUSION®

- Community
- Content
- Commerce

- Product information
- Product catalogs
- Stock information
- XML schemas
- Metatagging



DATAFUSION, Inc. 1999

DATAFUSION®

- 3000 years of library science.
- Infomediation and eCommerce.
- Controlled vocabularies.
- Solutions.

DATAFUSION, Inc. 1999

Five ways to organize things

DATAFUSION®

- Chronological
- Alphabetical
- Spatially
- Physical attributes (size, color, …)
- Topic

Richard Saul Wurman

DATAFUSION, Inc. 1999

What is a controlled vocabulary?

DATAFUSION®

A standard system of terminology used for coding, classifying, or otherwise uniquely identifying data and information.

- Glossaries
- Specialized dictionaries
- Standard terminology lists
- Reference data
- Authority files
- Classification schemes
- Domain-specific taxonomies
- Thesauri
- Ontologies

DATAFUSION, Inc. © 1999

Some aliases for Benzene

- Annulene
- Benzin
- Benzine
- Benzol
- Benzole
- Benzolene
- Bicarburet of Hydrogen
- Carbon oil
- Caswell No. 077
- CCRIS 70
- Coal naphtha
- Cyclohexatriene
- EINECS 200-753-7
- EPA Pesticide Chemical Code 008801
- HSDB 35
- Mineral naphtha
- Motor benzol
- NCI-C55276
- Nitration benzene
- NSC 67315
- Phene
- Phenyl Hydride
- Polystream
- Pyrobenzol
- Pyrobenzole

Source: *ChemName*

DATAFUSION, Inc. © 1999

What is the purpose of using a controlled vocabulary?

DATAFUSION®

Collect together information objects ...

- by the same creator,
- on the same topic,
- that are the same work,
- that are part of a series,

or that have other characteristics in common.

DATAFUSION, Inc. 1999

Authoritative schemes

Term	Aliases	Authority
AZ	Ariz. Arizona 85XXX	US Postal Service - abbreviations
IBM	International Business Machines Intl Bus Machines	NY Stock Exchange - ticker symbols
Master plans	General plans Comprehensive plans 27299	Art & Architecture Thesaurus - document types
nyctalopia	night blindness moon blindness WN1.6_NOUN:10438186	National Library of Medicine Medical Subject Headings - diseases
3571	Electronic computers	Standard Industrial Codes (SIC)
514191	Information Retrieval Services On-line Information Services	North American Industrial Classification System (NAICS)

DATAFUSION, Inc. © 1999

What is an ontology?

DATAFUSION®

- The branch of philosophy that deals with being.

  American Heritage Dictionary
- A taxonomy of everything that divides human knowledge or a subset of human knowledge into a clean set of categories, e.g., the Dewey Decimal System.

  http://fiat.gslis.utexas.edu/
- Formal, structured representations of a domain of knowledge …

  Murray. Technologies, Techniques, and Disciplines in Knowledge Management

DATAFUSION, Inc. 1999

What problems are you trying to solve?

DATAFUSION®

- Use and re-use existing information sources.
- Locate, gather, monitor and retrieve relevant information.
- Fuse content from disparate sources.
- Provide highly granular tagging.
- Fault-tolerant searching.
- Individualized presentation of results.

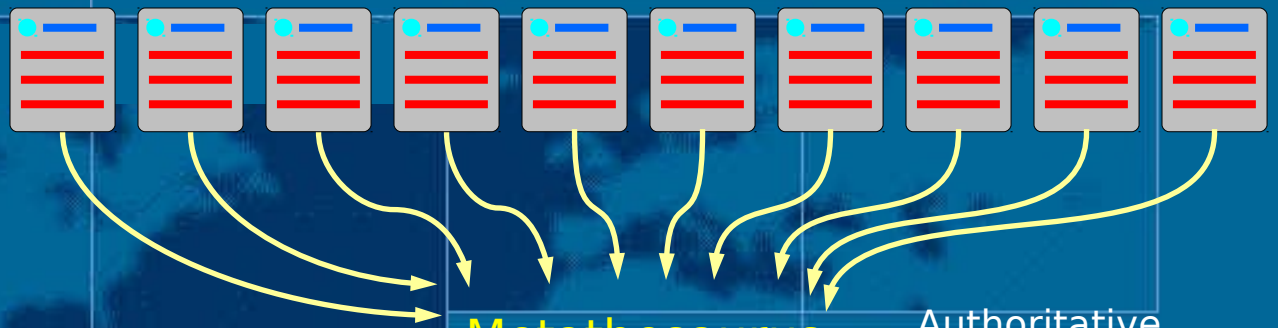
DATAFUSION, Inc. 1999

DATAFUSION®

- 3000 years of library science.
- Infomediation and eCommerce.
- Controlled vocabularies.
- Solutions.

DATAFUSION, Inc. 1999

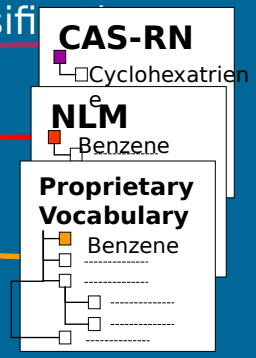
Content aggregation

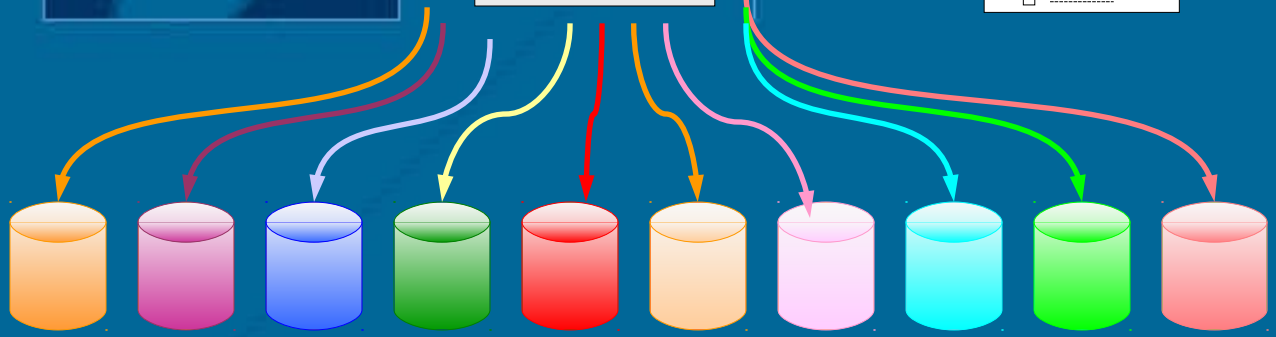


Source content

Metathesaurus

Authoritative Classifi

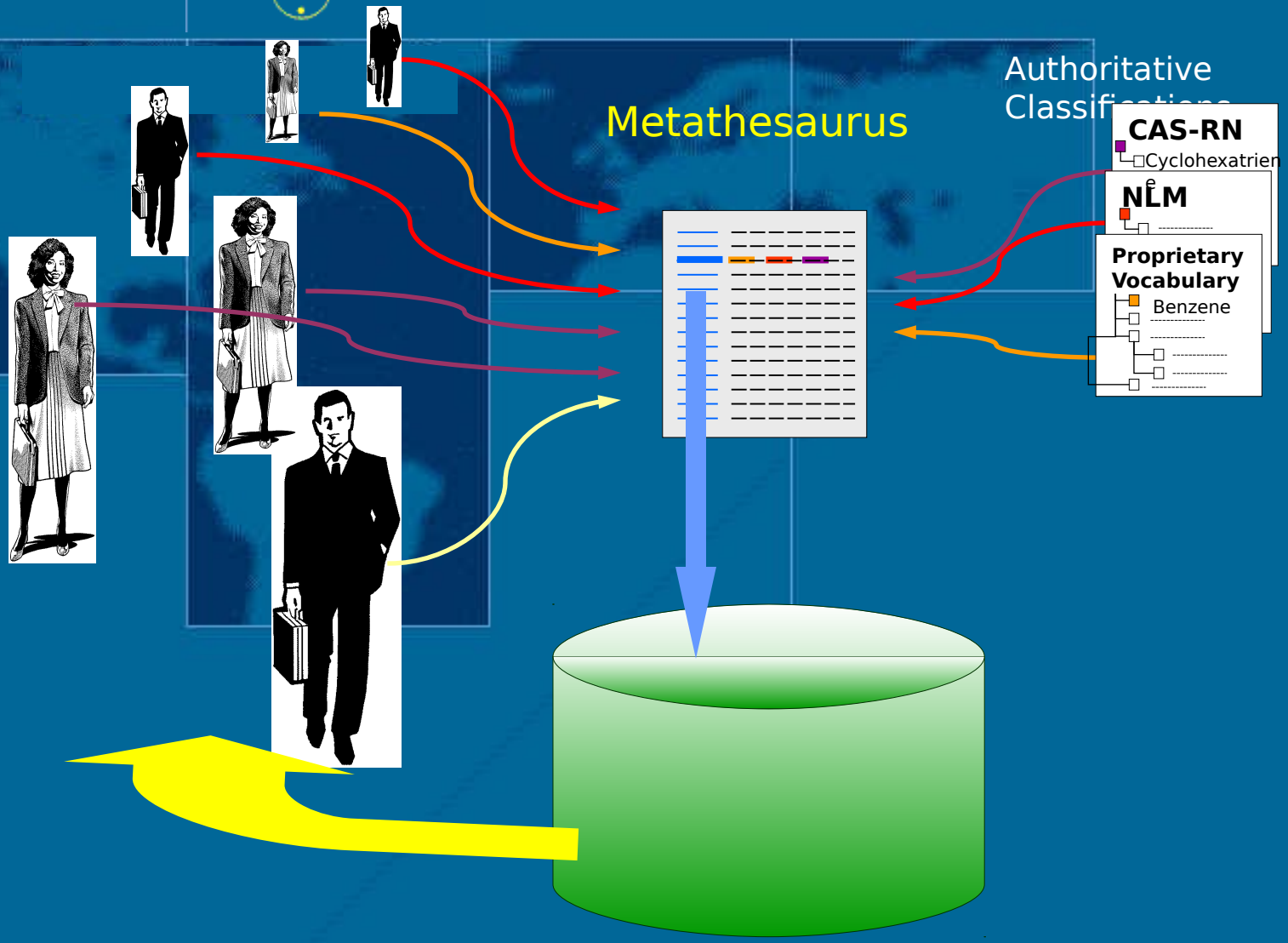




Custom Subsets

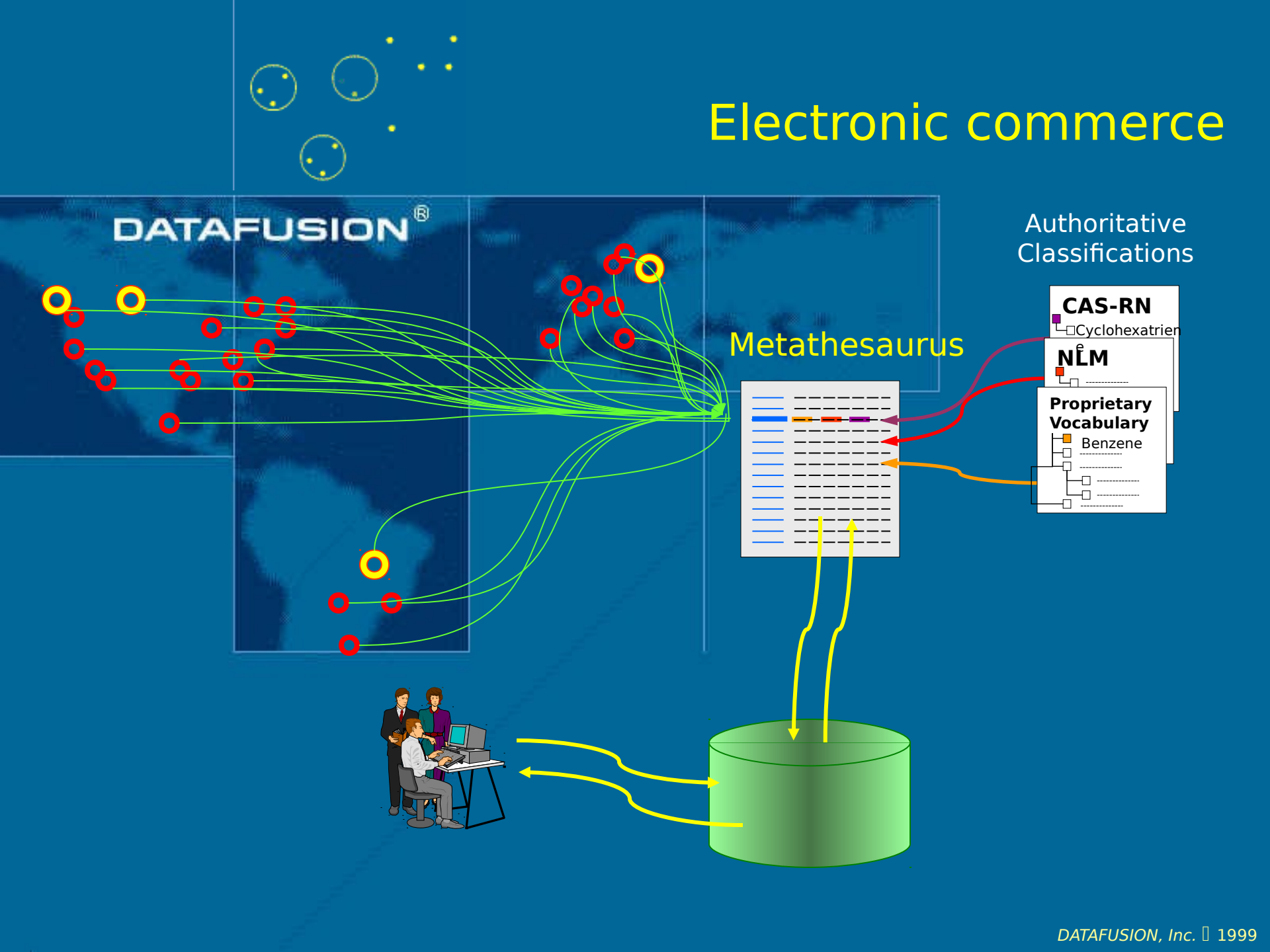
DATAFUSION, Inc. © 1999

Intelligent searching



DATAFUSION, Inc. © 1999

Electronic commerce



DATAFUSION, Inc. 1999

Summary

DATAFUSION®

- Information management is not a new problem.
- Library and information science methodologies and techniques still apply,

  especially controlled vocabularies.
- Operate at the metadata level, not on each information object itself.
- Take advantage of existing authorities.
- Semi-automated solutions work best.

DATAFUSION, Inc. © 1999

DATAFUSION®

Technology working with controlled vocabularies

Joseph A. Busch
DATAFUSION, Inc.
139 Townsend St.
San Francisco, CA 94110
(415) 222-0100
Jbusch@datafusion.net
http://www.datafusion.net/

DATAFUSION, Inc. 1999